Un modèle probabiliste pour la détection de l'incertitude dans le langage naturel

Pierre-Antoine Jean* — **Sébastien Harispe *** — **Sylvie Ranwez*** — **Patrice Bellot**** — **Jacky Montmain***

* *LGI2P, École des mines d'Alès, 69 rue Georges Besse F-30035 Nîmes cedex 1, {prenom.nom}@mines-ales.fr,*
** *LSIS, Avenue Escadrille Normandie-Niemen F-13397 Marseille cedex 20, patrice.bellot@lsis.org,*

RÉSUMÉ. *La détection de l'incertitude dans le langage naturel est centrale pour le développement de nombreux modèles exploitant l'analyse de textes e.g. questions-réponses, raisonnement approché, enrichissement de bases de connaissances. Après une synthèse des différentes classifications de l'incertitude et des méthodes de détection correspondantes, cet article introduit une approche supervisée et générique de détection de l'incertitude. Celle-ci se base sur l'analyse statistique de différentes caractéristiques lexicales et syntaxiques afin de construire une représentation vectorielle d'une phrase analysable par des méthodes de classification éprouvées. L'évaluation que nous proposons tient compte des différentes dimensions de l'incertitude et de la nature des textes. Les résultats obtenus sur différents jeux de validation soulignent la performance globale de la méthode proposée et ouvrent de nombreuses perspectives.*

ABSTRACT. *Designing approaches able to automatically detect uncertain natural language expressions is central to design efficient models based on text analysis – for domains such as question-answering, approximate reasoning, knowledge-based population. This article proposes an overview of several contributions and classifications defining the concept of uncertainty expressions in natural language, and their related detection methods that have been proposed so far. A new supervised and generic approach is next introduced for this specific task; it is based on the statistical analysis of multiples lexical and syntactic features used to characterize sentences through vector-based representations that can be analyzed by proven classification methods. The global performance of our approach is demonstrated and discussed with regard to various dimensions of uncertainty and text specificities.*

MOTS-CLÉS : *Détection de l'incertitude, Classification binaire, Modèle supervisé.*

KEYWORDS: *Uncertainty detection, Bianary classification, Supervised model.*

1. Introduction

Qu'elle soit d'ordre linguistique, numérique ou due à la subjectivité de certains jugements, l'incertitude est omniprésente dans toute situation langagière. En général levée par un *récepteur* humain qui réinterprète la phrase dans un contexte de co-énonciation (Fuchs, 2008), cette incertitude est beaucoup plus difficile à identifier de manière automatique dans des fragments de textes, qui peuvent de plus être sortis de leur contexte. Pourtant, ceux-ci peuvent être à la base d'un raisonnement approché ou, de façon plus globale, intégrés dans un processus décisionnel, pour ne citer que quelques-unes des applications possibles du traitement automatique des langues (TAL). La détection automatique de l'incertitude dans les textes a suscité un grand nombre de travaux ces dernières années et des événements majeurs comme la « *Conference on Natural Language Learning* » (CoNLL) en 2010 ont contribué au développement de méthodes dédiées. Leur intégration dans des applications d'analyse de sentiments, de recherche d'information, de questions-réponses ou encore d'extraction d'information à partir de textes a montré une réelle plus-value. Cependant, les formes diverses d'incertitude détectées ainsi que la forte dépendance de cette détection à la nature des textes analysés laissent largement ouvertes les perspectives de recherche dans ce domaine.

Caractériser l'incertitude renvoie rapidement à des dimensions diverses du texte. L'incertitude peut être interprétative par défaut ou excès de sens (présupposés, sous-entendus), ou bien provoquée par l'ambiguïté des termes (concurrence de sens, polysémie) (Fuchs, 2008). Plusieurs classifications de l'incertitude ont été proposées pour distinguer les différentes dimensions de l'incertitude et définir précisément celles qui peuvent être prises en compte dans un processus de détection automatique. (Jousselme *et al.*, 2003) présentent un état de l'art intéressant sur diverses classifications de l'incertitude. On y trouve notamment la classification de (Smets, 1997) proposée dans le domaine de la fusion de l'information, qui distingue l'incertitude comme une subdivision d'un concept plus général : l'imperfection de l'information. Si elle présente de façon claire les différentes formes d'incertitude, cette classification n'est cependant pas assez détaillée pour permettre une exploitation efficace dans la tâche de détection. Pour cela, on préférera la classification approfondie proposée par (Farkas *et al.*, 2010). Les auteurs y distinguent deux principales branches de l'incertitude : l'incertitude au niveau du discours et l'incertitude sémantique. L'**incertitude au niveau du discours** dénote dans la proposition du locuteur un manque d'information intentionnel ou non. Ainsi la proposition « Des personnes ont manifesté » appelle des compléments d'information : quelles personnes, combien étaient-elles ? (Ferson *et al.*, 2015). La subjectivité d'une proposition fait également partie de cette dimension de l'incertitude. Ainsi, l'incertitude au niveau du discours dépend principalement du contexte, du discours et de l'orateur ; en l'absence de connaissance sur ces différentes dimensions, l'incertitude persiste (Vincze, 2014). Par ailleurs, on appelle **incertitude sémantique** les propositions dont on ne peut pas déterminer la valeur de vérité étant donné l'état mental actuel du locuteur, soit le degré de confiance qu'il associe à sa proposition. Cette branche de l'incertitude se subdivise en deux catégories, d'une part l'incerti-

tude épistémique et d'autre part l'incertitude hypothétique. La principale différence entre ces deux catégories est que les propositions d'incertitude hypothétique peuvent être vraies, fausses ou incertaines *e.g. Il croit que la Terre est plate* (les connaissances actuelles du monde nous permettent d'infirmer cette proposition) tandis que les propositions d'incertitude épistémique sont définitivement incertaines *e.g. Il peut pleuvoir*, la factualité de la proposition ne peut être connue.

Cet intérêt pour l'incertitude est amplement justifié par le fait qu'elle est largement présente dans le langage naturel. (Light *et al.*, 2004) estiment que 11% des phrases dans les résumés des articles de MEDLINE sont incertaines. Toujours dans le domaine biomédical, le corpus de résumés BioScope (Szarvas *et al.*, 2008) contient 11871 phrases dont 2101 incertaines (17,5%), et si l'on s'intéresse à un domaine plus vaste, le corpus SFU (Konstantinova *et al.*, 2012) possède 17263 phrases tirées de documents dans divers domaines (films, livres, critiques) dont 23,7% incertaines.

Dans cet article, nous proposons une méthode de détection de l'incertitude basée sur une analyse statistique de différentes caractéristiques lexicales et syntaxiques. Cette méthode offre des résultats particulièrement intéressants lorsqu'elle est confrontée au processus de validation défini dans le cadre de la conférence CoNLL. Cette évaluation nous a permis différentes observations concernant, entre autres, l'influence de la nature des textes analysés et les dimensions de l'incertitude considérées. La section suivante analyse les différentes méthodes de détection d'incertitude citées dans la littérature, leurs caractéristiques et leurs performances. La section 3 détaille notre approche. Utilisant des techniques d'apprentissage automatique, cette approche matérialise les phrases comme des vecteurs de caractéristiques représentant des informations générales sur la phrase ou des informations basées sur les marqueurs lexicaux d'incertitude et leur contexte local. La section 4 présente les résultats de cette méthode confrontée aux jeux de tests de la conférence CoNLL 2010. Nous utilisons également des mesures éprouvées en classification de textes pour l'évaluer.

2. Les méthodes de détection de l'incertitude

De nombreux travaux, dans différents domaines, ont été consacrés à la détection des différentes formes d'incertitude et à leur prise en compte dans différentes applications de TAL, ce qui a permis d'en améliorer les performances. Par exemple, (Wu *et al.*, 2011) démontrent que la détection de l'incertitude permet d'améliorer la précision des informations extraites à partir de rapports radiologiques. Dans le domaine de l'analyse des sentiments, (Pang et Lee, 2004) ont montré que la détection de la subjectivité, considérée comme une forme d'incertitude au niveau du discours, aide à améliorer la classification de la polarité des phrases. En ce qui concerne les systèmes questions-réponses, (Ben Abacha, 2012) montre de manière empirique comment la détection de l'incertitude peut améliorer les performances du système MEANS.

L'incertitude s'exprime sous des formes diverses selon la nature des textes par l'emploi de verbes spéculatifs (suggérer, présumer), d'adjectifs et adverbes se rap-

portant naturellement à l'incertitude (probablement, possible), d'auxiliaires modaux permettant d'exprimer une modalité (pouvoir, devoir) ou encore l'emploi de certains temps ou modes de conjugaison (subjonctif, conditionnel).

Différentes approches ont été suggérées dans le domaine de la détection automatique de l'incertitude. Ces approches se focalisent soit sur une détection binaire de la certitude d'une phrase, soit sur la détection de la portée des marqueurs d'incertitude au sein de la phrase. Un défi proposé lors de la conférence CoNLL 2010 a notamment permis de confronter différentes approches pour ces deux tâches. L'évaluation des méthodes était réalisée au travers de deux corpus : BioScope et WikiWeasel (Farkas *et al.*, 2010). BioScope est un corpus spécifique au domaine biomédical alors que WikiWeasel est un corpus généraliste constitué de paragraphes de Wikipedia (*cf.* tableau 1). La principale différence entre ces deux corpus au regard de la classification de (Szarvas *et al.*, 2012) est le type d'incertitude considéré. BioScope prend en compte uniquement l'incertitude sémantique tandis que WikiWeasel ajoute à celle-ci la prise en compte d'une partie de l'incertitude au niveau du discours, notamment au travers des mots *weasel* qui se rapportent à la notion de source dans le texte : *Qui dit ça ?* et à la part de subjectivité apportée par un contributeur de Wikipedia. L'identification automatique de ces mots *weasel* a été étudiée par (Ganter et Strube, 2009).

BioScope	*We **suggest** that these IL-10 producing effector T cells **may** contribute to clearing malaria infection without-inducing immune-mediated pathology.*
WikiWeasel	*He was **probably** born in Spain, but **some sources** say he was born in Quito.*

Tableau 1. *Exemples de phrases issues du corpus BioScope et WikiWeasel. La phrase issue de BioScope révèle deux marqueurs d'incertitude épistémique et la phrase de WikiWeasel un marqueur d'incertitude épistémique et un autre d'incertitude au niveau du discours (manque d'information).*

L'approche ayant obtenu les meilleurs résultats pour la tâche de détection binaire sur le corpus BioScope (une F-mesure de 86,4%) a été proposée par (Tang *et al.*, 2010). Leur méthode se base sur trois classifieurs disposés en deux couches. La première couche comprend un CRF (*Conditional Random Fields*) et un SVM (*Support Vector Machine*) se basant tous les deux sur un même ensemble de caractéristiques (mot, lemme, préfixe, suffixe, morphosyntaxe, syntagme) et un système d'étiquettes identiques (BIO). La seconde couche, quant à elle, est constituée d'un autre CRF et utilise des caractéristiques provenant des résultats de la première couche. Cette dernière couche réalise la détection finale des marqueurs et chaque phrase contenant un marqueur est annotée comme incertaine.

Pour le corpus WikiWeasel, (Georgescul, 2010) a proposé la meilleure approche avec une F-mesure de 60,2% en utilisant une classification par SVM basée sur une fonction *kernel RBF* (*Radial Basis Function*). Une méthode similaire a été mise en place dans (Cruz *et al.*, 2015) et a obtenu une F-mesure de 92,3% sur le corpus SFU

dont les annotations suivent celles proposées dans BioScope. On peut d'ores et déjà remarquer la différence de performance entre les meilleures méthodes en fonction de la nature du corpus (très spécialisé ou généraliste) et des dimensions de l'incertitude considérées. D'autres méthodes intéressantes de détection binaire ont été proposées et appliquées sur WikiWeasel. Par exemple, (Chen et Eugenio, 2010) ont présenté une méthode hybride en deux phases. La première réalise une recherche par motif de mots consécutifs, dont certains sont généralisés par leur *morphosyntaxe* à l'aide de Lucene[1], pour récupérer des phrases candidates (potentiellement incertaines). La deuxième phase utilise ces phrases candidates comme entrées pour une classification par maximum d'entropie. Cette méthode a obtenu le troisième meilleur résultat sur 17 participants avec une F-mesure de 57,4%.

Les résultats obtenus à CoNLL entre les différents corpus nous dévoilent les limites des méthodes à être efficientes sur toutes les facettes de l'incertitude *e.g.* (Tang *et al.*, 2010) obtiennent 86,4% en F-mesure sur BioScope et 55% sur WikiWeasel correspondant à la meilleure moyenne (70.7%) de la conférence. Par conséquent, nous allons nous intéresser à la conception d'une méthode générique de détection de l'incertitude.

Dans la section suivante, nous présentons notre méthode basée sur une représentation vectorielle concise de la phrase – ce choix de représentation a été adopté afin d'éviter les biais de surapprentissage identifiés dans l'utilisation de représentations vectorielles de grandes tailles (Joachims, 2002) ; la taille réduite des vecteurs assure aussi une faible complexité de la tâche de classification, caractéristique souhaitée pour le traitement de gros volumes de données. La représentation vectorielle d'une phrase synthétise différentes statistiques propres à chaque caractéristique étudiée pour la détection d'incertitude (*e.g.* unigramme, bigramme). Plusieurs mesures fréquentistes sont ainsi proposées et évaluées pour le calcul de ces caractéristiques. Elles sont par la suite comparées aux mesures classiquement retrouvées dans la littérature associée à la classification de textes. Nous déléguons ensuite la tâche de classification basée sur l'analyse des représentations vectorielles à un SVM (Sebastiani, 2002).

3. Un modèle probabiliste pour la détection de l'incertitude

3.1. *Vue d'ensemble du modèle*

L'objectif est de distinguer si une phrase exprime de l'incertitude ou non – problématique de classification binaire. Pour cela, nous disposons d'un ensemble de phrases annotées S provenant des corpus BioScope, WikiWeasel ou SFU. De cet ensemble, il est possible d'extraire des informations sur les particularités lexicales et syntaxiques des phrases certaines et incertaines (*e.g.* la présence de marqueurs d'incertitude, les motifs morphosyntaxiques récurrents). Notre méthode propose de définir une représentation vectorielle sur un ensemble de caractéristiques d'une phrase. Chaque compo-

1. https //lucene.apache.org

sante du vecteur réalise une agrégation des poids affectés à une caractéristique locale dans la phrase (*e.g.* l'ensemble des unigrammes), en fonction d'une classe c à analyser (*e.g. est marqueur d'incertitude*). Cette représentation découle des méthodes de classification binaire de textes. En effet, dans le paradigme des méthodes d'apprentissage automatique, une des principales problématiques de la classification de textes est la manière de représenter un document. Celui-ci est généralement matérialisé comme un vecteur de poids associés à ses différentes caractéristiques pouvant être les mots d'un vocabulaire dans les approches les plus simples (Sebastiani, 2002). Ces poids ont pour objectif de sélectionner les caractéristiques les plus pertinentes d'une classe c afin de réduire l'espace des dimensions associé à l'ensemble des caractéristiques d'un corpus (Yang et Pedersen, 1997).

La sous-section suivante présente les différentes caractéristiques utilisées par notre méthode, et leur utilisation dans un modèle d'apprentissage supervisé. Les modalités d'évaluation et les résultats obtenus seront discutés dans la section 4.

3.2. *Définition des caractéristiques locales et globales*

Les fonctions caractéristiques sélectionnées et étudiées, bien que générales et se voulant indépendantes d'un domaine particulier, traduisent l'intuition que certains marqueurs lexicaux et sémantiques semblent importants pour la classification d'une phrase. Deux niveaux de granularité ont été considérés pour la définition de ces fonctions.

Le premier niveau s'intéresse aux spécificités globales d'une phrase traduisant potentiellement une expression d'incertitude *e.g.* nous considérons la taille d'une phrase car nous supposons que la longueur est un indice discriminant.

Le second niveau porte sur les motifs *n-grammes* qui composent une phrase. Nous entendons par motifs *n-grammes* les séquences de *n* éléments de même nature, par exemple, la forme lemmatisée des mots ou leur symbole morphosyntaxique (*PoS*). A chaque *n-gramme* est associé un poids exprimant le fait qu'il puisse traduire une expression d'incertitude. La composante de la projection de la phrase selon la caractéristique analysée tiendra compte de l'agrégation des poids associés aux *n-grammes* qui la composent. Chacune des caractéristiques est ainsi définie par un quadruplet (type, taille, contexte, agrégation) précisant le type de *n-gramme* analysé (lemme et motif morphosyntaxique), la taille des *n-grammes* (n), le contexte, i.e. si le score des *n-grammes* se base sur la fréquence d'observations des *n-grammes* dans une phrase étiquetée incertaine ou comme marqueur explicite d'incertitude (ces derniers sont précisés dans le jeu de d'entraînement), et l'agrégation utilisée pour résumer les scores des différents *n-grammes* de la phrase pour cette caractéristique. Le tableau 2 résume les différentes caractéristiques basées sur l'analyse de *n-grammes*. La figure 1 détaille le calcul de deux caractéristiques. Une phrase est donc caractérisée par un vecteur de $\mathbb{R}^6$ avant l'étape de classification – les cinq caractéristiques présentées dans le tableau 2 et la taille de la phrase (F_5).

	Type	Taille	Contexte	agrégation
F_1	Lemme	1	Marqueur d'incertitude	somme
F_2	Lemme	2	Marqueur d'incertitude	somme
F_3	Lemme	1	$\in$ à une phrase incertaine	somme
F_4	*PoS*	5	$\in$ à une phrase incertaine	somme
F_6	Lemme	1	$\in$ à une phrase incertaine	max

Tableau 2. *Description des caractéristiques locales utilisées.*

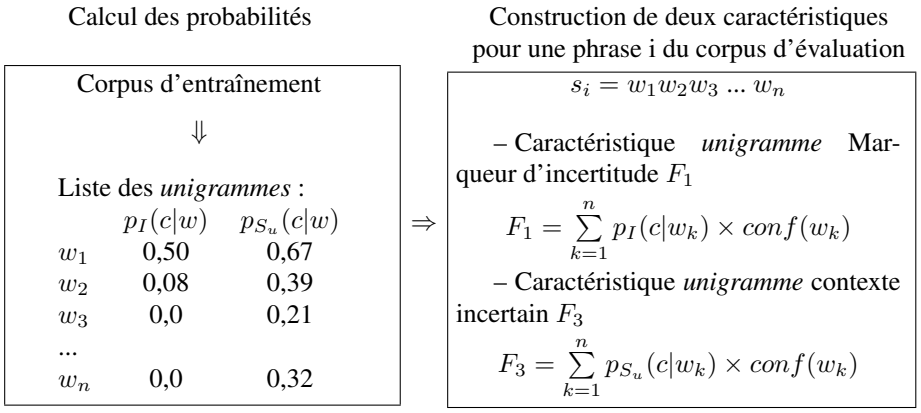

Figure 1. *Calcul de deux caractéristiques d'une phrase s_i appartenant au corpus d'évaluation. La première phase (à gauche) permet de calculer les probabilités sur le corpus d'entraînement avec $p_I(c|w)$ la probabilité conditionnelle qu'un lemme w soit marqueur d'incertitude et $p_{S_u}(c|w)$ la probabilité conditionnelle qu'un lemme w soit présent dans une phrase incertaine. Ainsi, la signification de la classe c dépend de la probabilité conditionnelle considérée. La deuxième phase correspond à la construction de deux caractéristiques F_1 et F_3 définies dans le tableau 2. Le score de confiance $conf$ est détaillé dans la sous-section 3.3.*

La sous-section suivante présente les différentes mesures étudiées afin de calculer le score (poids) associé à chaque motif de *n-gramme*. Par commodité, nous illustrons désormais nos propos en considérant les observations en tant que marqueurs d'incertitude pour un lemme. Par conséquent, les exemples ne vont pas tenir compte du type *PoS*, du type lemme avec une taille supérieure à 1 et du contexte *appartient à une phrase incertaine* (*cf.* tableau 2). Nous partons également du postulat qu'un marqueur d'incertitude traduit une phrase incertaine.

3.3. *Définition d'une mesure probabiliste*

Les données d'entraînement définissent un ensemble de phrases incertaines $S_u \subset S$ avec S l'ensemble des phrases. Ces données nous permettent d'obtenir pour chaque lemme w son nombre d'occurrences dans le corpus, noté $\#_S(w)$, son nombre d'occurrences dans les phrases incertaines, $\#_{S_u}(w)$ avec $\#_{S_u}(w) \leq \#_S(w)$ ainsi, que son nombre d'occurrences en tant que marqueur d'incertitude $\#_{I_{S_u}}(w)$, avec I_{S_u} l'ensemble des marqueurs d'incertitude du corpus et $\#_{I_{S_u}}(w) \leq \#_{S_u}(w)$. Connaissant le lemme w, nous pouvons alors définir la probabilité conditionnelle qu'il soit marqueur d'incertitude *i.e.* qu'il appartienne à la classe c (*cf.* équation 1). La signification de la classe c et la définition de cette probabilité dépendent du contexte de la caractéristique considérée (marqueur d'incertitude ou *appartient à une phrase incertaine – cf.* tableau 2).

$$p_I(c|w) = \#_{I_{S_u}}(w)/\#_S(w) \quad [1]$$

Cependant, l'analyse de cette probabilité dans le but de distinguer les marqueurs d'incertitude n'est pas suffisante. Du fait de la taille limitée des corpus d'entraînement, il est en effet fréquent d'obtenir des probabilités très élevées pour certains termes, malgré leur présence limitée dans le corpus d'entraînement. Prenons le cas extrême d'un lemme w qui n'apparaît qu'une seule fois dans le corpus et ce, de façon fortuite, dans un contexte incertain, sa probabilité d'appartenir à une phrase incertaine serait alors : $p_I(c|w) = 1$.

Afin de pallier cette limite, nous définissons un score de *confiance* associé à cette probabilité qui évalue la pertinence de considérer le motif analysé (ici le lemme w) comme marqueur d'incertitude. Dans la modélisation de ce score de confiance, nous cherchons à considérer à la fois le nombre d'occurrences $\#_S(w)$ et la probabilité $p(c)$ qu'un lemme, observé dans l'ensemble des mots du corpus W et tiré aléatoirement, soit marqueur d'incertitude (*cf.* équation 2). Par conséquent, si un lemme obtient une forte probabilité d'être marqueur d'incertitude, la confiance dans ce score sera d'autant plus élevée que ce lemme est représentatif du corpus et que la probabilité $p(c)$ est faible.

$$p(c) = \frac{\sum_{w \in W} \#_{I_{S_u}}(w)}{\sum_{w \in W} \#_S(w)} \quad [2]$$

Pour la modélisation de ce score de confiance, nous avons étudié deux mesures, possédant une sémantique propre, utilisant comme paramètre $\#_S(w)$ et $p(c)$ ainsi qu'une mesure témoin utilisant uniquement $\#_S(w)$. Le premier score de confiance étudié repose sur une loi de distribution binomiale cumulée utilisant : $p(c)$, la probabilité de tirer un marqueur d'incertitude dans W, le nombre d'occurrences $\#_{I_{S_u}}(w)$ du mot w observé en tant que marqueur d'incertitude et le nombre d'occurrences $\#_S(w)$ du mot w dans le corpus complet. Cette loi est définie par la probabilité de fonction de masse suivante, avec $n = \#_S(w)$, $k = \#_{I_{S_u}}(w)$ et $p = p(c)$:

$$p_b(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \quad [3]$$

Ainsi, la confiance associée à la probabilité conditionnelle $p_I(c|w)$ est fonction de la probabilité d'effectuer un nombre d'observations identifiées comme marqueurs d'incertitude supérieur ou égal à $\#_{I_{S_u}}(w)$ (loi cumulative) en effectuant $\#_S(w)$ tirages aléatoires. Par conséquent, plus la valeur associée à la loi binomiale cumulative est élevée et moins $p_I(c|w)$ traduit une incertitude. Le score de confiance est alors modélisé par $1 - p_b(X \geq k)$.

La seconde modélisation de la confiance que nous avons étudiée suppose intuitivement que plus la probabilité $p(c)$ est grande, plus le nombre d'occurrences d'un motif doit être conséquent pour associer un score de confiance élevé à la probabilité qu'il soit marqueur d'incertitude. Cette représentation de la confiance peut être modélisée par une fonction sigmoïde de $\#_S(w)$ dont le paramètre $p(c)$ caractérise la courbure. Plus $p(c)$ est grand et plus la pente de la courbe est lissée. (*cf.* figure 2).

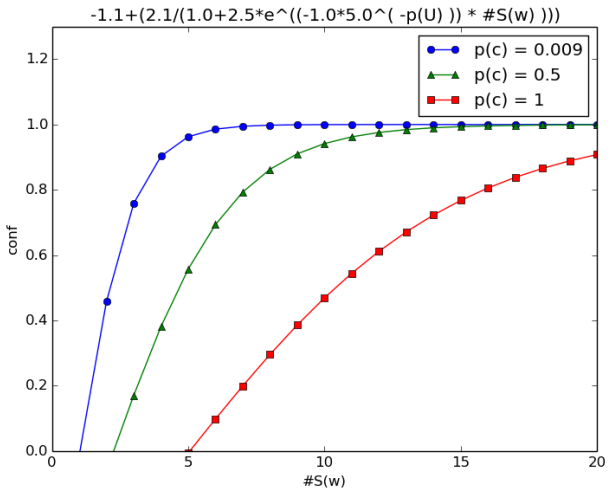

Figure 2. *Modélisation de la confiance en fonction du paramètre $p(c)$. La courbe avec les ronds a un $p(c) = 0.009$ correspondant à la probabilité pour un mot d'être marqueur d'incertitude dans le corpus BioScope. La courbe avec les triangles a un $p(c) = 0.5$ et la courbe avec les carrés un $p(c) = 1$.*

Finalement, la dernière modélisation de la confiance tient uniquement compte de $\#_S(w)$. Cette mesure, nous permet d'observer l'impact de la probabilité $p(c)$ dans les modélisations précédentes. Elle signifie que plus un lemme est fréquent, plus la confiance qui lui sera accordée sera importante (*cf.* équation 4).

$$conf(w) = 1 - \frac{1}{\#_S(w)} \quad [4]$$

La fonction F_1 utilisée pour calculer une des dimensions de la représentation vectorielle d'une phrase s, qui caractérise les unigrammes marqueurs d'incertitude est fonction de : la probabilité que l'unigramme traduise une forme d'incertitude, modulée par $conf(w)$ la confiance associée à la probabilité de l'unigramme w (*cf.* équation 5). Cette formulation est généralisable à l'ensemble des caractéristiques.

$$F_1(s) = \sum_{k=1}^{n} p_I(c|w_k) \times conf(w_k) \quad [5]$$

3.4. *Sélection automatique des caractéristiques optimales*

Les vecteurs représentant les phrases sont par la suite utilisés comme entrée d'un modèle d'apprentissage automatique SVM[2]. Les natures très différentes des corpus BioScope, WikiWeasel et SFU font qu'ils n'ont pas le même ensemble de caractéristiques optimales dans le SVM (*cf.* tableau 3). Ainsi, une stratégie de sélection automatique des caractéristiques optimales à partir d'un corpus d'entraînement a été appliquée en suivant les travaux de (Chen et Lin, 2006). Ces travaux mettent en avant l'utilisation d'une forêt aléatoire. Ainsi, nous avons mis en place une procédure de sélection récursive où les caractéristiques en dessous d'un certain pourcentage d'importance sont supprimées jusqu'à obtenir les caractéristiques les plus pertinentes.

Caractéristiques	BioScope	WikiWeasel	SFU
F_1 - Unigramme, marqueurs d'incertitude	x	x	x
F_2 - Bigramme, marqueurs d'incertitude		x	
F_3 - Unigramme, dans les phrases incertaines	x	x	x
F_4 - Motifs *PoS* taille 5, dans les phrases incertaines		x	
F_5 - $\|s\|$ la taille de la phrase s		x	
F_6 - max(Unigramme marqueurs d'incertitude)	x	x	x
F_7 - Trigramme, marqueurs d'incertitude			

Tableau 3. *Les caractéristiques optimales pour BioScope, WikiWeasel et SFU obtenues par l'application d'une forêt aléatoire. La caractéristique des trigrammes n'est jamais pertinente quel que soit le jeu de données.*

Les travaux de (Øvrelid *et al.*, 2010) suggèrent que les caractéristiques syntaxiques ne sont pas nécessaires dans la tâche de détection de l'incertitude. Cependant, on

2. Ce modèle se base sur une fonction *kernel* RBF (Gaspar *et al.*, 2012) et est optimisé au niveau des paramètres C et γ selon l'étude réalisée par (Georgescul, 2010)

remarque que pour WikiWeasel la caractéristique des motifs *PoS* contenus dans les phrases incertaines est discriminante.

4. Résultats et discussion

Cette section présente les résultats obtenus en utilisant la probabilité conditionnelle $p_I(c|w)$ couplée avec les différentes définitions de la confiance présentées dans la section précédente : la loi binomiale cumulée, la fonction sigmoïde et la confiance témoin ($1-1/\#_S(w)$). Les résultats sont ensuite comparés en modifiant la probabilité $p_I(c|w)$ par des mesures éprouvées en théorie de l'information.

4.1. *Résultats de l'approche probabiliste*

Le tableau 4 précise les résultats obtenus en utilisant la probabilité d'être marqueur d'incertitude couplée aux différents scores de confiance définis précédemment. Chaque entrée donne la précision, le rappel et la F-mesure obtenus pour chaque expérience.

	BioScope	WikiWeasel	SFU corpus
Loi binomiale	77,9 / 82,9 / **80,3**	66,7 / 25 / 36,3	87,8 / 95,8 / 91,6
Sigmoïde	75,8 / 82,1 / 78,8	73,8 / 43,6 / 54,8	88,2 / 96,6 / **92,2**
$1 - 1/\#_S(w)$	75,8 / 81,6 / 78,6	64,9 / 48,8 / **55,7**	88,2 / 96,4 / 92,1
Sans confiance	76,3 / 81,2 / 78,7	72,1 / 11,8 / 20,2	88,1 / 95,9 / 91,9

Tableau 4. *Résultats de la méthode en utilisant différentes confiances associées à la probabilité $p_I(c|w)$ sur les corpus BioScope, WikiWeasel et SFU. Les caractéristiques utilisées entre les différentes confiances sont fixées selon le jeu de données.*

Ces résultats amènent plusieurs interprétations. Dans un premier temps, on remarque que les différents scores de confiance n'impactent pas ou peu les résultats sur les jeux de données BioScope et SFU, considérant tous les deux uniquement l'incertitude sémantique. Cependant, la confiance améliore les scores sur WikiWeasel. Une analyse fine de ces corpus nous permet d'observer que la disparité des marqueurs d'incertitude pour WikiWeasel est bien plus grande que pour BioScope, elle-même plus grande que celle de SFU. Cette forte disparité, représentée par le nombre, la nature et la distribution des marqueurs d'incertitude à l'échelle du corpus ajoute un bruit important dans ces données. Pour diminuer ce bruit, nous avons appliqué un filtre de pré-classification sur le nombre d'occurrences $\#_{S_u}(w)$ des motifs *n-grammes* dans les phrases incertaines (*cf.* équation 6).

$$\#_{S_u}(w) = \begin{cases} \#_{S_u}(w) & \text{si } \#_{I_{S_u}}(w) \geq 1 \\ 0 & \text{sinon.} \end{cases} \quad [6]$$

Ce filtre impacte les caractéristiques considérant les lemmes et les motifs morpho-syntaxiques dans le contexte des phrases incertaines. Les résultats en appliquant ce filtre sont indiqués dans le tableau 5.

Dans un second temps, on s'aperçoit que les résultats pour la loi binomiale sur WikiWeasel sont moins bons qu'avec les autres scores de confiance. Ces résultats s'expliquent par la faible valeur de la probabilité de succès $p(c)$ (*e.g.* 0,03 pour WikiWeasel) qui a pour effet d'augmenter le score de confiance de la loi binomiale.

Enfin, les résultats sur les différents corpus d'évaluation soulignent l'importance de la définition de l'incertitude que l'on souhaite détecter et de la nature des textes.

	BioScope	WikiWeasel	SFU corpus
Loi binomiale	76,2 / 82,9 / 79,4	61,3 / 61,5 / 61,4	88,1 / 96,7 / 92,2
Sigmoïde	76,3 / 82,9 / **79,5**	69,7 / 55,3 / 61,7	88,3 / 96,7 / **92,3**
$1 - 1/\#_S(w)$	76,3 / 82,9 / **79,5**	68,9 / 57,7 / **62,8**	88,3 / 96,7 / **92,3**
Sans confiance	76,2 / 82,9 / 79,4	64,7 / 54,5 / 59,2	88,3 / 96,7 / **92,3**

Tableau 5. *Résultats de la méthode en utilisant un filtre sur le nombre d'occurrences des lemmes présents dans le contexte des phrases incertaines. Les caractéristiques utilisées entre les différentes confiances sont fixées selon le jeu de données.*

Les résultats du tableau 5 démontrent l'efficacité du filtre sur les résultats de WikiWeasel en éliminant la majeure partie du bruit issu des motifs présents dans les phrases incertaines. Finalement, ce filtre permet de compenser la faiblesse de la méthode lorsqu'elle est appliquée sur des jeux de données dont la disparité des marqueurs est forte.

Ces résultats améliorent ceux de l'approche de (Georgescul, 2010) sur le corpus WikiWeasel, qui avait obtenu la première place de la tâche 1 lors de CoNLL 2010 sur ce même corpus avec une F-mesure de 60,2%. De plus, nous obtenons la meilleure moyenne en terme de F-mesure, 71,2%, sur les jeux BioScope et WikiWeasel par rapport à la meilleure moyenne de la conférence, 70,7% par (Tang *et al.*, 2010). Au niveau du corpus de SFU, non-utilisé dans CoNLL 2010, nous avons des résultats similaires à ceux de (Cruz *et al.*, 2015).

4.2. *Comparaison avec d'autres mesures*

L'utilisation de la probabilité conditionnelle $p_I(c|w)$ a été confrontée à des mesures couramment utilisées dans le domaine de la classification de textes. Ces métriques considèrent un lemme w et sa relation avec une classe c. L'ensemble des valeurs obtenues à ces différents tests sont données dans le tableau 6.

Pointwise mutual information, PMI, mesure l'association d'un lemme w avec la classe c (*cf.* équation 7). Cette mesure est proche de la définition de notre probabilité

$p_I(c|w)$. Elle pondère simplement cette probabilité par $p(c)$. Cependant, cette probabilité $p(c)$ est très faible lorsqu'on considère la classe *est marqueur d'incertitude* et aura pour conséquence de bruiter la valeur de la probabilité.

$$pmi(w,c) = log(\frac{p(c,w)}{p(c).p(w)}) = log(\frac{p(c|w)}{p(c)}) \quad [7]$$

Odds Ratio mesure le degré de dépendance entre un lemme w et la classe c (*cf.* équation 8). Appliqué à nos données, le *Odds Ratio* favorise les motifs avec un faible écart entre $\#_{I_{S_u}}(w)$ et $\#_S(w)$.

$$orr(w,c) = log(\frac{p(w|c).(1-p(w|\bar{c}))}{p(w|\bar{c}).(1-p(w|c))}) \quad [8]$$

Categorical Proportional Difference, CPD, est un ratio qui considère pour un lemme w le nombre de documents appartenant aux classes c et $\bar{c}$ qui le contiennent. L'équation 9 définit CPD avec dw_c le nombre de documents de la classe c contenant w, $dw_{\bar{c}}$ le nombre de documents de la classe $\bar{c}$ contenant w. Dans notre problématique de détection binaire de l'incertitude au niveau de la phrase, cette mesure a été adaptée, dw_c représente le nombre d'occurrences du lemme w en tant que marqueur d'incertitude.

$$cpd(w,c) = \frac{dw_c - dw_{\bar{c}}}{dw} \quad [9]$$

Weighted Log Likelihood Ratio mesure la dissimilarité de la distribution du lemme w en fonction des classes c et $\bar{c}$ (*cf.* équation 10).

$$wllr(w,c) = p(w|c).log(\frac{p(w|c)}{p(w|\bar{c})}) \quad [10]$$

Nous avons couplé ces différentes métriques avec les mesures de confiance définies dans la sous-section 3.3. Ce couplage s'apparente à l'adaptation de modèles classiquement retrouvés pour la classification de textes. Par exemple, (Hamdan, 2015) définit le poids final d'un terme par la formule $w_i = localWeight \times globalWeight \times normalization$ avec $localWeight$ une mesure fréquentiste du terme dans le document (*e.g.* $log(termFrequency + 1)$), $globalWeight$ une métrique appliquée aux termes à l'échelle du corpus (présentée en début de section) et $normalization$ permet d'ajuster les poids en fonction de la taille du document. Les résultats sont présentés dans le tableau 6. Une analyse a également été menée en amont sur ces mesures ainsi que sur les mesures suivantes : *Chi Square*, *Natural Entropy* et *Kullback-Leibler Divergence*. Cette étude a porté sur l'analyse du comportement de chaque mesure par rapport à la contrainte principale fixée pour notre modèle. Nous l'avons vu, cette contrainte repose sur la prise en compte, lorsque la probabilité conditionnelle est fixe, du nombre d'observations $\#_S(w)$ pour le calcul du score, tel que pour deux mots w_1

Métrique	Confiance	BioScope	WikiWeasel	SFU
PMI	$log(\#_S(w))$	75,6%	33,8%	88,3%
	$1 - 1/\#_S(w)$	77,3%	40,6%	91,1%
	Loi binomiale	76,6%	52,3%	91,5%
	Sigmoïde	77,1%	37,7%	91%
	Sans confiance	76,4%	35,1%	90,6%
Odds Ratio	$log(\#_S(w))$	78,1%	45,5%	91,1%
	$1 - 1/\#_S(w)$	**79,3%**	52%	92,1%
	Loi binomiale	**79,3%**	**55%**	**92,2%**
	Sigmoïde	**79,3%**	51,5%	92,1%
	Sans confiance	79,2%	51,3%	92,1%
CPM	$log(\#_S(w))$	70,8%	45,2%	78,6%
	$1 - 1/\#_S(w)$	70,4%	49,9%	78%
	Loi binomiale	69,7%	48,1%	80,1%
	Sigmoïde	70,5%	48,6%	78,1%
	Sans confiance	69,6%	48%	73,3%
Wllr	$log(\#_S(w))$	53,7%	16,5%	69,8%
	$1 - 1/\#_S(w)$	55,1%	11%	66,3%
	Loi binomiale	55,5%	45%	67,1%
	Sigmoïde	55,1%	11,6%	65,8%
	Sans confiance	55,1%	18,9%	65,7%

Tableau 6. *F-mesure des confiances associées à différentes métriques globales étudiées sur les corpus BioScope, WikiWeasel et SFU. Les caractéristiques utilisées entre les différentes confiances sont fixées selon le jeu de données. Le filtre sur le nombre d'occurrences des lemmes présents dans les phrases incertaines est appliqué.*

et w_2 avec $\#_S(w_1) > \#_S(w_2)$ et $p_I(c|w_1) = p_I(c|w_2)$ le score de la mesure soit supérieur pour $w1$.

Dans la problématique de détection de l'incertitude, la probabilité conditionnelle $p_I(c|w)$ obtient de meilleurs résultats comparée aux autres métriques globales. De plus, cette probabilité couplée avec la confiance $1 - 1/\#_S(w)$ est la plus performante en moyenne sur les jeux de données.

5. Conclusion et Perspectives

Dans cet article, nous avons proposé une méthode d'apprentissage automatique pour la détection binaire de l'incertitude dans le langage naturel. Cette méthode se base sur une représentation vectorielle concise ($\mathbb{R}^6$) de la phrase construite à partir des différentes probabilités conditionnelles de motifs *n-grammes* pondérées par un score

de confiance fréquentiste. L'approche obtient des résultats intéressants au regard de toutes les dimensions de l'incertitude.

Plusieurs pistes d'amélioration de la méthode sont envisagées. Ces pistes concernent notamment le calcul des poids des motifs *n-grammes*. En effet, un mécanisme de propagation basé sur l'analyse des collocations permettrait une pondération contextuelle plus précise (Lavalley *et al.*, 2010) ; ceci dans le but d'éviter des erreurs de classification dues au poids d'un lemme trop discriminant. Une autre piste d'amélioration serait d'ajouter une caractéristique contextuelle au niveau de la phrase *i.e.* indiquer par une valeur booléenne si la phrase précédente est détectée comme incertaine. On considère dans ce cas l'hypothèse qu'une phrase aura plus de chance d'être incertaine si les phrases précédentes sont incertaines. Une autre amélioration serait d'étendre la nature des motifs utilisés dans les caractéristiques. Actuellement, seulement deux types sont utilisés, les lemmes et les motifs morphosyntaxiques. Nous pourrions par exemple expérimenter les étiquettes d'un arbre des dépendances en tant qu'unité de base d'un motif ou élaborer un motif hybride de plusieurs types (Chen et Eugenio, 2010). Enfin, nous envisageons d'étendre les comparaisons effectuées dans cette étude en prenant notamment en compte des travaux récents dans le domaine des réseaux neuronaux démontrant des performances intéressantes dans la tâche de classification de la polarité des phrases (Tai *et al.*, 2015).

6. Bibliographie

Ben Abacha A., « Recherche de réponses précises à des questions médicales : le système de questions-réponses MEANS », *PhD, Univ Paris Sud-Paris XI,* p. 162, 2012.

Chen L., Eugenio B. D., « A Lucene and Maximum-Entropy Model based hedge detection system », *Fourteenth Conference on Computational Natural Language Learning,* p. 114-119, 2010.

Chen Y. W., Lin C. J., « Combining SVMs with various feature selection strategies », *Feature extraction,* p. 315-324, 2006.

Cruz N., Taboada M., Mitkov R., « A machine learning approach to negation and speculation detection », *Association for Information Science and Technology*, 2015.

Farkas R., Vincze V., Móra G., Csirik J., Szarvas G., « The CoNLL-2010 shared task : learning to detect hedges and their scope in natural language text », *Fourteenth Conference on Computational Natural Language Learning,* p. 1-12, 2010.

Ferson S., O'Rawe J., Antonenko A., Siegrist J., Mickley J., Luhmann C. C., Sentz K., Finkel A. M., « Natural language of uncertainty : numeric hedge words », *International Journal of Approximate Reasoning*, vol. 57, p. 19-39, 2015.

Fuchs C., « L'incertitude interprétative dans l'activité de langage », *revue de l'IUF*, vol. 5, p. 41-57, 2008.

Ganter V., Strube M., « Finding hedges by chasing weasels : Hedge detection using Wikipedia tags and shallow linguistic features », *ACL-IJCNLP*, 2009.

Gaspar P., Carbonell J., Oliveira J. L., « On the parameter optimization of Support Vector Machines for binary classification », *J Integr Bioinform*, vol. 9, n° 3, p. 201, 2012.

Georgescul M., « A Hedgehop over a Max-Margin Framework Using Hedge Cues », *Fourteenth Conference on Computational Natural Language Learning,* p. 26-31, 2010.

Hamdan H., Sentiment Analysis in Social Media, P.h.d thesis, Université d'Aix-Marseille, 2015.

Joachims T., « Learning to classify text using support vector machines : Methods, theory and algorithms », *Kluwer Academic Publishers*, vol. , p. 205, 2002.

Jousselme A. L., Maupin P., Bosse E., « Uncertainty in a situation analysis perspective », *Sixth International Conference of Information Fusion,* p. 1207-1214, 2003.

Konstantinova N., de Sousa S. C., Díaz N. P. C., López M. J. M., Taboada M., Mitkov R., « A review corpus annotated for negation, speculation and their scope », *LRE,* p. 3190-3195, 2012.

Lavalley R., Clavel C., Bellot P., « Extraction probabiliste de chaînes de mots relatives à une opinion », *Traitement Automatique des Langues*, vol. 51, p. 101-130, 2010.

Light M., Qiu X. Y., Srinivasan P., « The language of bioscience : Facts, speculations, and statements in between », *BioLink 2004 workshop on linking biological literature, ontologies and databases : tools for users,* p. 17-24, 2004.

Pang B., Lee L., « A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts », *The 42nd annual meeting on Association for Computational Linguistics*, vol. Association for Computational Linguistics, p. 271, 2004.

Sebastiani F., « Machine learning in automated text categorization », *ACM computing surveys*, vol. 34, n° 1, p. 1-47, 2002.

Smets P., « Imperfect information : Imprecision and uncertainty », *Uncertainty Management in Information Systems* p. 225-254, 1997.

Szarvas G., Vincze V., Farkas R., Csirik J., « The BioScope corpus : annotation for negation, uncertainty and their scope in biomedical texts », *Workshop on Current Trends in Biomedical Natural Language Processing,* p. 38-45, 2008.

Szarvas G., Vincze V., Farkas R., Móra G., Gurevych I., « Cross-genre and cross-domain detection of semantic uncertainty », *Computational Linguistics*, vol. 38, n° 2, p. 335-367, 2012.

Tai K. S., Socher R., Manning C. D., « Improved semantic representations from tree-structured long short-term memory networks », *eprint arXiv :1503.00075*, 2015.

Tang B., Wang X., Wang X., Yuan B., Fan S., « A Cascade Method for Detecting Hedges and their Scope in Natural Language Text », *Fourteenth Conference on Computational Natural Language Learning,* p. 13-17, 2010.

Vincze V., « Uncertainty Detection in Natural Language Texts », *PhD, University of Szeged,* p. 141, 2014.

Wu A. S., Do B. H., Kim J., Rubin D. L., « Evalution of negation and Uncertainty detection and its impact on precision and recall in search », *Journal of Digital Imaging*, vol. 24, n° 2, p. 234-242, 2011.

Yang Y., Pedersen J. O., « A comparative study on feature selection in text categorization », *ICML*, vol. 97, p. 412-420, 1997.

Øvrelid L., Velldal E., Oepen S., « Syntactic scope resolution in uncertainty analysis », *23rd International Conference on Computational Linguistics*, vol. 10, p. 1379-1387, 2010.